
Biométrie

# Étude de la stabilité de l'ACP par la méthode du Bootstrap

C. Baril
CIRAD-Forêt

Septembre, 1993

# Étude de la stabilité de l'ACP par la méthode du Bootstrap

Biométrie

C. Baril
CIRAD-Forêt

Septembre, 1993

# SOMMAIRE

# ETUDE DE LA STABILITE DE L'ACP PAR LA METHODE DU BOOTSTRAP

## INTRODUCTION

Les observations étant toujours sujettes aux erreurs de mesure ou d'échantillonnage, les résultats d'une Analyse en Composantes Principales (ACP) peuvent varier d'un échantillon à l'autre.

La méthode de ré-échantillonnage du Bootstrap permet, grâce à la simulation de plusieurs échantillons, d'estimer la variabilité des résultats de l'ACP et donc d'estimer leur fiabilité.

## 1. RAPPELS

### a/ But de l'ACP

Le but de l'ACP est de rechercher une approximation de la matrice de données initiale X(n,p), à n individus et p variables mesurées sur chaque individu, par une matrice de rang inférieur q. Si les n lignes de X sont considérées comme les coordonnées de n points dans un espace à p dimensions, on peut alors représenter graphiquement X dans un sous-espace de plus faible dimension q.

Donc, le problème qui se pose concerne le choix du nombre de composantes (ou de dimensions) qui doivent être retenues. Les règles souvent utilisées prennent en compte deux types de critères :

- la part de variance expliquée (ou taux d'inertie),
- le comportement en "éboulis" des valeurs-propres.

### b/ Principe de l'ACP

L'ajustement progressif de sous-espaces "emboîtés" de dimensions k croissantes (k = 1, ... ,p) au nuage de n points, par la méthode des moindres carrés, fournit les p vecteurs-propres ($u_k$) correspondant aux p valeurs-propres ($\lambda_k$) de la matrice symétrique X'X, (avec $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$). Les coordonnées $Z_k$ des points-individus sur l'axe k sont alors les produits scalaires constituant les lignes de $X\, u_k$

On écrit : $$\underset{(p,n)}{X'}\ \underset{(n,p)}{X}\ \underset{(p,p)}{U} = \underset{(p,p)}{U}\ \underset{(p,p)}{\Lambda}$$

et l'on obtient la matrice des variables transformées par l'ACP :

$$\underset{(n,p)}{Z} = \underset{(n,p)}{X}\ \underset{(p,p)}{U}$$

On note :
- $V = X'X$ la matrice de variance-covariance de X
- $\Delta = diag(V)$ la matrice diagonale des variances de X
- $\Lambda$ = la matrice diagonale des valeurs-propres de X'X
- U = la matrice orthogonale des vecteurs-propres de X'X
- R = la matrice des corrélations de X

## 2. APPLICATION DU BOOTSTRAP A L'ACP

### a/ Principe du Bootstrap

La méthode du Bootstrap consiste à réaliser un tirage avec remise de B n-sous-échantillons à partir d'un n-échantillon, lui-même représentatif d'une population d'origine (P). Si l'on s'intéresse à une statistique T, la distribution F de T du n-échantillon dans la population d'origine P est simulée par la distribution $F^{(b)}$ de la même statistique $T^{(b)}$ (b = 1, 2, ... ,B) calculée dans chaque sous-échantillon. Cette simulation donne accès aux caractéristiques de $F^{(b)}$ (la moyenne, le biais, la variance...) qui sont les estimations des caractéristiques de F.

### b/ Tirage des sous-échantillons

Le $b^{ème}$ n-sous-échantillon est obtenu par tirage équiprobable, avec remise, de n unités dans le n-échantillon.

Soit **X** la matrice des observations dans l'échantillon,

soit $P^{(b)}$ une matrice des pondérations associée au $b^{ème}$ sous-échantillon, de terme général $p_{ij}$

on peut noter la matrice des observations du $b^{ème}$ sous-échantillon :

$$\underset{(n,p)}{X^{(b)}} = \underset{(n,n)}{P^{(b)}}\ \underset{(n,p)}{X}$$

**Exemple** :

Prenons $n = 4$ et admettons que l'on tire :

- 2 fois la $1^{ère}$ ligne d'observations dans **X**
- 1 fois la $2^{ème}$ ligne d'observations dans **X**
- 0 fois la $3^{ème}$ ligne d'observations dans **X**
- 1 fois la $4^{ème}$ ligne d'observations dans **X**

On aura : $$\mathbf{P}^{(b)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

avec :
$$\begin{cases} \sum_i p_{ij} = \text{nombre d'occurences de la ligne } j \\ \sum_j p_{ij} = 1 \\ \sum_{ij} p_{ij} = n \end{cases}$$

Le Bootstrap peut être réalisé pour l'ACP non-normée (diagonalisation de la matrice de variance-covariance V) ou normée (diagonalisation de la matrice des corrélations R).

Trois natures de tirages sont envisageables pour chaque type d'ACP :

- des tirages sur les données brutes,
- des tirages sur les données centrées,
- des tirages sur les variables transformées.

Dans tous les cas, les données "bootstrappées" constituant chaque sous-échantillon s'écriront sous la forme du produit d'une matrice de pondérations par la matrice de données associée à l'échantillon considéré.

c/ Notations

Des exposants permettent de signifier l'origine des paramètres :

- paramètres associés à la population de base : exposant (0),
- paramètres associés au n-échantillon : pas d'exposant,
- paramètres associés aux n-sous-échantillons : exposant (b).

Pour préciser la nature des données "bootstrappées" et le type de l'ACP étudié, les paramètres sont indicés :

- ACP non-normée : indice (V),
- ACP normée : indice (R).

## 3. EVALUATION DE LA STABILITE DE L'ACP

• Pour l'ACP non-normée, les paramètres intéressants sont :

$$V^{(0)},\ \Lambda_V^{(0)} \text{ et } U_V^{(0)}$$

leurs estimations sur l'échantillon sont respectivement :

$$V,\ \Lambda_V \text{ et } U_V$$

La stabilité de l'ACP est mesurée par la variabilité de ces paramètres :

$$E\left(V - V^{(0)}\right)^2 \quad , \quad E\left(\Lambda_V - \Lambda_V^{(0)}\right)^2 \quad \text{et} \quad E\left(U_V - U_V^{(0)}\right)^2$$

La méthode du Bootstrap fournit une estimation de cette variabilité par :

$$E_B\left(V^{(b)} - V\right)^2 \quad , \quad E_B\left(\Lambda_V^{(b)} - \Lambda_V\right)^2 \quad \text{et} \quad E_B\left(U_V^{(b)} - U_V\right)^2$$

où $E_B$ est l'espérance calculée sur l'ensemble des B sous-échantillons.

• Pour l'ACP normée, les paramètres intéressants sont :

$$R^{(0)},\ \Lambda_R^{(0)} \text{ et } U_R^{(0)}$$

La stabilité de l'ACP est mesurée par :

$$E\left(R - R^{(0)}\right)^2 \quad , \quad E\left(\Lambda_R - \Lambda_R^{(0)}\right)^2 \quad \text{et} \quad E\left(U_R - U_R^{(0)}\right)^2$$

et estimée par :

$$E_B\left(R^{(b)} - R\right)^2 \quad , \quad E_B\left(\Lambda_R^{(b)} - \Lambda_R\right)^2 \quad \text{et} \quad E_B\left(U_R^{(b)} - U_R\right)^2$$

Quel que soit le type d'ACP, afin d'attribuer un scalaire à chacune de ces mesures, on "déroule" chaque matrice (p,p) en un vecteur ($p^2$,1) avant de calculer les espérances.

**a/ Etude de la stabilité de l'ACP non-normée**

La méthode du Bootstrap peut être appliquée indifféremment aux données brutes (X) ou aux données centrées (Y = X - $\mu$). En effet, la variance étant invariante par translation :

$$V_Y = V_X \text{ et } V_Y^{(b)} = V_X^{(b)}$$

donc $E_B\left(V_Y^{(b)} - V_Y\right)^2$ peut être utilisée pour estimer $E\left(V_X - V_X^{(0)}\right)^2$

Une troisième possibilité consiste en un tirage de B sous-échantillons parmi les variables transformées : Z

où :
$$\underset{(n,p)}{Z} = \underset{(n,p)}{X}\ \underset{(p,p)}{U_{XV}}$$

En fait, cette procédure est équivalente aux précédentes. En effet :

$$\left.\begin{array}{l} \bullet\ V_Z^{(b)} = U_{XV}'\ V_X^{(b)}\ U_{XV} \\ \quad V_Z = U_{XV}'\ V_X\ U_{XV} \end{array}\right\} \text{ donc } V_X^{(b)} - V_X = U_{XV}\left(V_Z^{(b)} - V_Z\right)U_{XV}'$$

et par conséquent $E_B\left(U_{XV}\left(V_Z^{(b)} - V_Z\right)U_{XV}'\right)^2$ estime $E\left(V_X - V_X^{(0)}\right)^2$

$$\left.\begin{array}{l} \bullet\ V_Z^{(b)} = U_{XV}'\ V_X^{(b)}\ U_{XV} \\ \quad V_X^{(b)} = U_{XV}^{(b)}\ \Lambda_X^{(b)}\ U_{XV}'^{(b)} \end{array}\right\} \text{ donc } V_Z^{(b)} = U_{XV}'\ U_{XV}^{(b)}\ \Lambda_X^{(b)}\ U_{XV}'^{(b)}\ U_{XV}$$

avec $V_Z^{(b)} = U_{ZV}^{(b)}\ \Lambda_Z^{(b)}\ U_{ZV}'^{(b)}$

$$\left.\begin{array}{l} \text{on obtient } \Lambda_X^{(b)} = \Lambda_Z^{(b)} \\ \text{et } U_{ZV}'^{(b)} = U_{XV}'^{(b)} U_{XV} \end{array}\right\} \text{donc} \left\{\begin{array}{l} E_B\left(\Lambda_Z^{(b)} - \Lambda_X\right)^2 \text{ estime } E\left(\Lambda_X - \Lambda_X^{(0)}\right)^2 \\ E_B\left(U_{ZV}'^{(b)} U_{XV}' - U_{XV}'\right)^2 \text{ estime } E\left(U_{XV}' - U_{XV}'^{(0)}\right)^2 \end{array}\right.$$

Travailler sur Z présente deux intérêts :

- la différence entre $V_Z^{(b)}$ et $V_Z$ est facile à visualiser car $V_Z$ est diagonale (en effet, $V_Z$ n'est autre que $\Lambda_{XV}$),
- la diagonalisation de $V_Z^{(b)}$ est plus économique en temps de calcul.

De plus, lorsque le nombre de variables est élevé, la seule prise en compte des k premières variables de l'ACP peut permettre de réduire davantage le temps de calcul tout en fournissant une bonne approximation de $V_X$ et $V_X^{(b)}$

**b/ Etude de la stabilité de l'ACP normée**

Dans le cas de l'ACP normée, la normalisation fournie par $\Delta_X^{-1/2}$ est différente pour chaque sous-échantillon. La matrice des corrélations s'écrit :

$$R_X^{(b)} = \Delta_X^{(b)-1/2} \; V_X^{(b)} \; \Delta_X^{(b)-1/2}$$

De même que lors de l'étude de la stabilité de l'ACP non-normée, on peut indifféremment utiliser **X**, **Y** ou **W** (avec $W = Y \Delta_X^{-1/2}$, la matrice des données centrées réduites) dans le calcul de $R_X^{(b)}$

*On appelle* [A] *la méthode du Bootstrap qui consiste à mesurer la stabilité des estimations à partir de* $R_X^{(b)}$ (ou $R_Y^{(b)}$ ou $R_W^{(b)}$)

Par contre, à l'inverse des résultats obtenus pour l'ACP non-normée, il devient sans intérêt d'utiliser la matrice des corrélations $R_S^{(b)}$ des variables transformées par l'ACP dans les sous-échantillons : **S**

où : $$\underset{(n,p)}{S} = \underset{(n,p)}{W} \; \underset{(p,p)}{U_{XR}}$$

En effet, la matrice des corrélations correspondantes s'écrit :

$$R_S^{(b)} = \left[\mathrm{diag}\left(U_{XR}' \; V_W^{(b)} \; U_{XR}\right)\right]^{-1/2} U_{XR}' \; V_W^{(b)} \; U_{XR} \left[\mathrm{diag}\left(U_{XR} \; V_W^{(b)} \; U_{XR}'\right)\right]^{-1/2}$$

Or, il est impossible de retrouver $R_X^{(b)}$ à partir de $R_S^{(b)}$ et donc d'estimer la stabilité de l'ACP normée en appliquant la méthode du Bootstrap aux variables transformées. Toutefois, à défaut de pouvoir estimer $R_X^{(b)}$, on peut s'en approcher en estimant $V_W^{(b)}$

*On appelle* [B] *la méthode du Bootstrap qui consiste à mesurer la stabilité des estimations à partir de* $V_W^{(b)}$. *Cette approximation se justifie par le fait qu'il est raisonnable de penser que le comportement de* $R_X - R_X^{(0)}$ *n'est pas trop éloigné de celui de* $V_W - R_X^{(0)}$

$$\left.\begin{array}{l} \text{Soit} \quad V_W^{(b)} = U_{XR} \; V_S^{(b)} \; U_{XR}' \\ \text{et} \quad\;\; R_X \;\;\; = U_{XR} \; \Lambda_{XR} \; U_{XR}' \end{array}\right\} \quad \text{donc} \quad V_W^{(b)} - R_X = U_{XR} \left(V_S^{(b)} - \Lambda_{XR}\right) U_{XR}'$$

par conséquent $E_B\left(U_{XR}\left(V_S^{(b)} - \Lambda_{XR}\right) U_{XR}'\right)^2$ estime $E\left(V_W - R_X^{(0)}\right)^2$

## 4. MESURES DE LA STABILITE DE L'ACP

Afin de calculer un nombre raisonnable de paramètres, il est préférable de se concentrer sur les plus intéressants. En Analyse en Composantes Principales, le problème est de pouvoir déterminer la dimension de l'espace de représentation optimum. Il s'agit de conserver toutes les caractéristiques stables et importantes des données étudiées tout en ignorant les axes instables et sans signification.

Dans cet esprit, on peut envisager de mesurer la stabilité de deux types de paramètres :

- les valeurs-propres,
- les sous-espaces de représentation.

### a/ Stabilité des valeurs-propres

Soit $\Lambda$, la matrice diagonale des valeurs-propres ordonnées $(\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p)$, pour chaque axe k, on calcule la moyenne $(\bar{\lambda}_k)$, l'écart-type $(S_{\lambda k})$ et le biais associés à la valeur-propre $\lambda_k$ correspondante.

On écrit :

$$\begin{cases} \bar{\lambda}_k = \dfrac{1}{B}\sum_b \lambda_k^{(b)} \\ S_{\lambda k} = \sqrt{\dfrac{1}{B}\sum_b \left(\lambda_k^{(b)} - \bar{\lambda}_k\right)^2} \\ \text{Biais}_{\lambda k} = E_B\left(\lambda_k^{(b)} - \lambda_k\right) = \bar{\lambda}_k - \lambda_k \end{cases}$$

On démontre que le biais associé à la première valeur-propre est positif, tandis que le biais associé à la dernière valeur-propre est négatif. En d'autres termes, la proportion de variabilité expliquée par le $1^{er}$ axe de l'ACP sur un sous-échantillon ne peut être que supérieure ou égale à la proportion de variabilité expliquée par le $1^{er}$ axe de l'ACP réalisée sur l'échantillon d'origine.

Par construction, le raisonnement inverse doit être appliqué à la proportion de variabilité expliquée par le dernier axe de l'ACP appliquée à un sous-échantillon de l'échantillon d'origine.

Soit $Z_1, \ldots, Z_p$ les variables de l'ACP, on obtient par conséquent :

$$\begin{cases} \text{var}\left(Z_1^{(0)}\right) \leqslant \text{var}\left(Z_1\right) = \lambda_1 & \text{donc} \quad E_B\left(\lambda_1^{(b)} - \lambda_1\right) \geqslant 0 \\ \text{var}\left(Z_p^{(0)}\right) \geqslant \text{var}\left(Z_p\right) = \lambda_p & \text{donc} \quad E_B\left(\lambda_p^{(b)} - \lambda_p\right) \leqslant 0 \end{cases}$$

**b/ Stabilité des sous-espaces de représentation**

Soit $E_k$, l'espace engendré par les k premiers vecteurs-propres, on mesure la variabilité de $E_k$ autour de $E_k^{(0)}$ par la mesure angulaire entre sous-espaces utilisée en analyse canonique :

$$A_k = \sum_{i,j \leqslant k} \rho_{ij}^2 \qquad \text{avec} \qquad \rho_{ij} = \text{corr}\left(Z_i^{(0)}, Z_j\right)$$

où $\rho_{ij}$ est la corrélation entre la $i^{ème}$ variable de l'ACP sur la population et la $j^{ème}$ variable de l'ACP sur l'échantillon.

**Remarque** : *la stabilité parfaite de tous les* $k^{èmes}$ *axes est caractérisée par :*

$$E\left(\rho_{kk}^2\right) = 1 \qquad \text{pour } k = 1, \ldots, p$$

*ce qui revient à écrire :* $E(A_k) = k \quad \forall\, k$

Pour chaque axe k, on calcule la moyenne $\left(\bar{A}_k\right)$, l'écart-type $(S_{Ak})$ et l'écart quadratique moyen $(E\ Q\ M_k)$ par rapport à k.

On écrit :

$$\begin{cases} \bar{A}_k = \dfrac{1}{B}\displaystyle\sum_b A_k^{(b)} \\ S_{Ak} = \sqrt{\dfrac{1}{B}\displaystyle\sum_b \left(A_k^{(b)} - \bar{A}_k\right)^2} \\ EQM_k = E_B\left(A_k^{(b)} - k\right)^2 \end{cases}$$

Les $A_k^{(b)}$ sont obtenus à partir des $\rho_{ij}^{(b)}$ mesurés de la façon suivante :

$$\left.\begin{array}{l} \bullet\ Z_j^{(b)} = X\ U_j^{(b)} \\ \quad \text{Var}\left(Z_j^{(b)}\right) = U_j^{'(b)} V\ U_j^{(b)} \\ \bullet\ Z_i = X\ U_i \\ \quad \text{Var}\left(Z_i\right) = U_i'\ V\ U_i = \lambda_i \\ \bullet\ Z_j^{'(b)} Z_i = U_j^{'(b)}\ V\ U_i \end{array}\right\} \text{donc } \rho_{ij} = \left(U_j^{'(b)}\ V\ U_i\right)\lambda_i^{-\frac{1}{2}}\left(U_j^{'(b)}\ V\ U_j^{(b)}\right)^{-\frac{1}{2}}$$
$$= \left(\lambda_i^{\frac{1}{2}}\ U_j^{'(b)} U_i\right)\left(U_j^{'(b)}\ V\ U_j^{(b)}\right)^{-\frac{1}{2}}$$

où V est la matrice de corrélation $R_X$ ou la matrice de variance-covariance $V_X$, selon que l'ACP est normée ou non.

$U_i$ est le $i^{ème}$ vecteur-propre de V.

## 5. EXEMPLES

La stabilité de l'ACP normée est ici étudiée sur deux séries de données (Cf. Daudin *et al.*, 1988) selon les critères définis au paragraphe précédent.

- Dans le premier jeu de données, 8 variables caractérisent la composition de 85 conteneurs de lait.

- Dans le second jeu de données, la prise de poids hebdomadaire de 28 chèvres est mesurée durant 11 semaines, la première variable étant le poids noté la première semaine.

Pour ces deux exemples, 100 sous-échantillons ont été tirés, par la méthode du Bootstrap, à partir de l'échantillon d'origine.

### a/ Composition du lait

*Résultats*

L'ACP du premier jeu de données se distingue par un premier axe expliquant 72 % de la variabilité (Cf. figure 1). Les critères de stabilité mesurés pour chacune des méthodes de Bootstrap [A] et [B], présentées dans le paragraphe 3, sont indiqués dans le tableau 1. Enfin, l'évolution des $A_k^{(b)}$ (avec b = 1,..., 100 et k = 1,..., 8) est représentée sur la figure 2 pour la méthode [A] et sur la figure 3 pour la méthode [B].

*Interprétation*

Il faut tout d'abord souligner que la stabilité parfaite du sous-espace de dimension p est naturelle car $E_p$ est engendré par l'ensemble des variables.

Il semble que l'on puisse tenir compte des quatre premiers axes sans risquer qu'un nouvel échantillonnage sur la même population n'aboutisse à des résultats différents.

On doit cependant noter que les valeurs-propres associées aux deux ou trois premiers axes ont une variance élevée, ce qui tendrait à prouver que si leur direction est stable, leur part d'explication de la variabilité globale est sujette à caution.

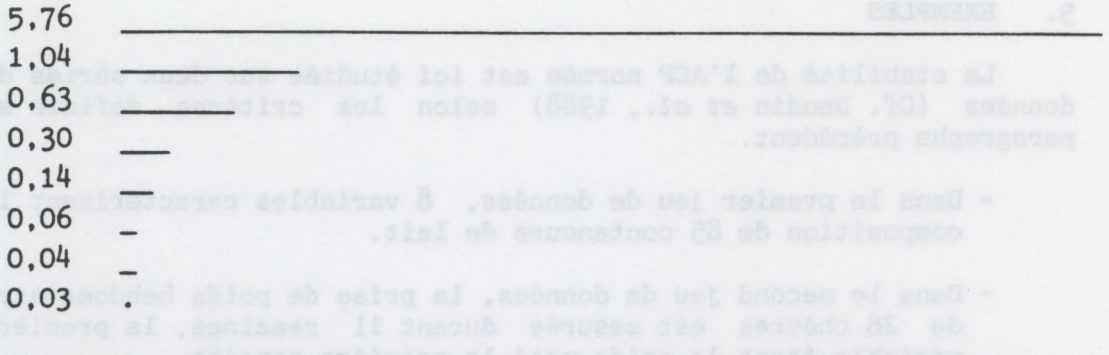


Figure 1 - *Valeurs-propres de l'ACP sur l'échantillon des données de composition du lait*

Tableau 1 - Critères de stabilité des valeurs-propres et des sous-espaces de représentation mesurés sur les données de composition du lait (calculés pour 100 sous-échantillons)

| Dimension k du sous-esp. | Méthode [A] | | | | | | Méthode [B] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{\lambda}_k$ | $S_{\lambda k}$ | Biais | $\bar{A}_k$ | $S_{Ak}$ | $EQM_k$ | $\bar{\lambda}_k$ | $S_{\lambda k}$ | $\bar{A}_k$ | $S_{Ak}$ | $EQM_k$ |
| 1 | 5,76 | 0,29 | 0,00 | 1,00 | 0,00 | 0,00 | 5,74 | 1,19 | 1,00 | 0,00 | 0,00 |
| 2 | 1,07 | 0,14 | 0,03 | 1,96 | 0,05 | 0,06 | 1,02 | 0,12 | 1,96 | 0,07 | 0,08 |
| 3 | 0,63 | 0,12 | 0,00 | 2,99 | 0,02 | 0,02 | 0,60 | 0,09 | 2,98 | 0,03 | 0,04 |
| 4 | 0,30 | 0,06 | 0,00 | 3,99 | 0,03 | 0,03 | 0,28 | 0,04 | 3,98 | 0,03 | 0,04 |
| 5 | 0,14 | 0,04 | 0,00 | 4,96 | 0,12 | 0,13 | 0,14 | 0,03 | 4,99 | 0,03 | 0,03 |
| 6 | 0,06 | 0,02 | 0,00 | 5,76 | 0,29 | 0,38 | 0,06 | 0,02 | 5,76 | 0,03 | 0,24 |
| 7 | 0,03 | 0,01 | -0,01 | 6,74 | 0,24 | 0,35 | 0,03 | 0,01 | 6,77 | 0,02 | 0,23 |
| 8 | 0,02 | 0,01 | -0,01 | 8,00 | 0,00 | 0,00 | 0,02 | 0,01 | 8,00 | 0,00 | 0,00 |

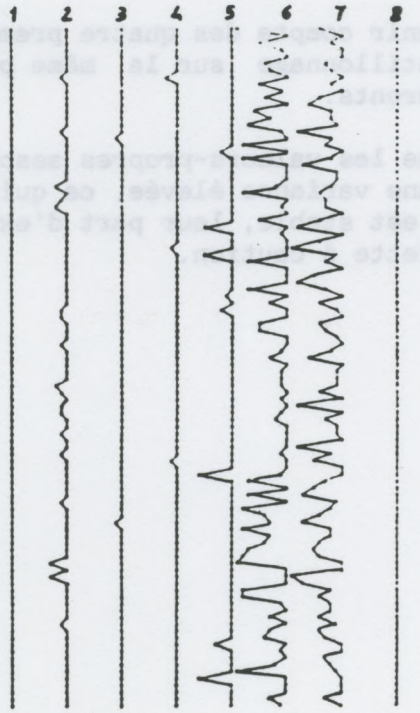


Fig. 2. Graphical representation of $A_i^{(b)}$ for 100 subsamples of the milk composition data. (Method [*A*]). The value of $A_i^{(b)}$ is given along the *x*-axes, the number of the subsample, *B* is given in the *y*-axes.

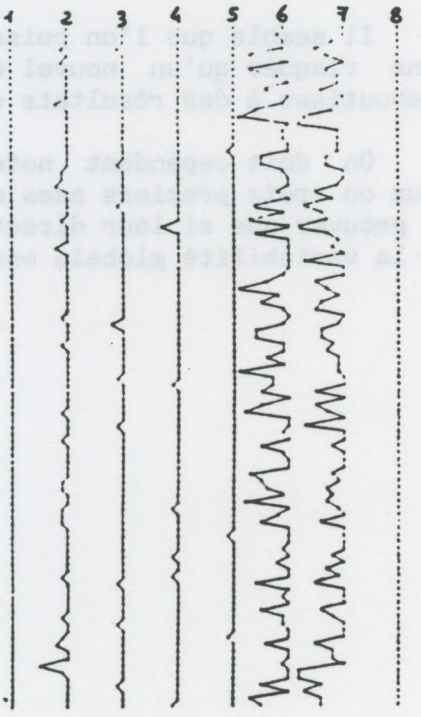


Fig. 3. Graphical representation of $A_i^{(b)}$ for 100 subsamples of the milk composition data. (Method [*B*]). The value of $A_i^{(b)}$ is given along the *x*-axes, the number of the subsample, *B* is given in the *y*-axes.

**b/ Accroissement du poids des chèvres**

*Résultats*

Les valeurs-propres de l'ACP du second jeu de données sont indiquées sur la figure 4. Les critères de stabilité mesurés pour les deux méthodes de Bootstrap [A] et [B] sont présentés dans le tableau 2. Comme les résultats obtenus pour ces deux méthodes sont similaires, seule l'évolution des $A_k^{(b)}$ pour la méthode [B] est visualisée (figure 5).

*Interprétation*

L'interprétation des résultats des simulations réalisées à partir du second lot de données est malheureusement sans appel. Le seul sous-espace relativement stable est le sous-espace de dimension 10 ($EQM_{10}$ = 0,05).

Ceci conduit l'utilisateur à n'apporter aucun crédit à toute représentation graphique de ces données. On peut, par exemple, imaginer que l'expérimentation a été largement perturbée par la présence d'individus malades.

Dans un tel contexte, la seule utilité de l'ACP se limite à la détection de données aberrantes. Toute conclusion concernant les variables serait hasardeuse.

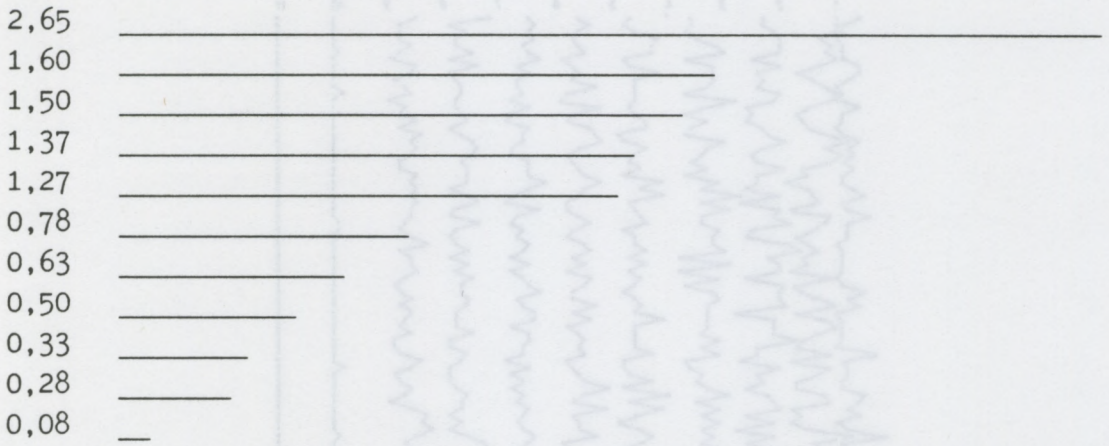


Figure 4 - *Valeurs-propres de l'ACP sur l'échantillon des données d'accroissement du poids des chèvres*

Tableau 2 - Critères de stabilité des valeurs-propres et des sous-espaces de représentation mesurés sur les données d'accroissement du poids des chèvres (calculés pour 100 sous-échantillons)

| Dimension k du sous-esp. | Méthode [A] | | | | | | Méthode [B] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{\lambda}_k$ | $S_{\lambda k}$ | Biais | $\bar{A}_k$ | $S_{Ak}$ | $EQM_k$ | $\bar{\lambda}_k$ | $S_{\lambda k}$ | $\bar{A}_k$ | $S_{Ak}$ | $EQM_k$ |
| 1 | 3,02 | 0,37 | 0,37 | 0,78 | 0,25 | 0,33 | 3,06 | 0,66 | 0,77 | 0,23 | 0,32 |
| 2 | 2,12 | 0,25 | 0,52 | 1,35 | 0,33 | 0,73 | 2,00 | 0,32 | 1,35 | 0,31 | 0,72 |
| 3 | 1,64 | 0,16 | 0,14 | 2,21 | 0,30 | 0,84 | 1,50 | 0,22 | 2,22 | 0,28 | 0,83 |
| 4 | 1,29 | 0,15 | -0,08 | 3,33 | 0,26 | 0,72 | 1,15 | 0,19 | 3,31 | 0,25 | 0,73 |
| 5 | 0,99 | 0,14 | -0,28 | 4,61 | 0,22 | 0,45 | 0,85 | 0,16 | 4,56 | 0,21 | 0,49 |
| 6 | 0,71 | 0,12 | -0,07 | 5,53 | 0,23 | 0,52 | 0,60 | 0,12 | 5,55 | 0,20 | 0,49 |
| 7 | 0,50 | 0,10 | -0,13 | 6,61 | 0,19 | 0,43 | 0,42 | 0,08 | 6,60 | 0,15 | 0,43 |
| 8 | 0,34 | 0,08 | -0,16 | 7,65 | 0,19 | 0,40 | 0,29 | 0,08 | 7,69 | 0,17 | 0,35 |
| 9 | 0,22 | 0,06 | -0,11 | 8,64 | 0,19 | 0,41 | 0,18 | 0,05 | 8,66 | 0,16 | 0,38 |
| 10 | 0,12 | 0,05 | -0,16 | 9,97 | 0,04 | 0,05 | 0,11 | 0,04 | 9,97 | 0,04 | 0,05 |
| 11 | 0,04 | 0,02 | -0,04 | 11,00 | 0,00 | 0,00 | 0,03 | 0,02 | 11,00 | 0,00 | 0,00 |

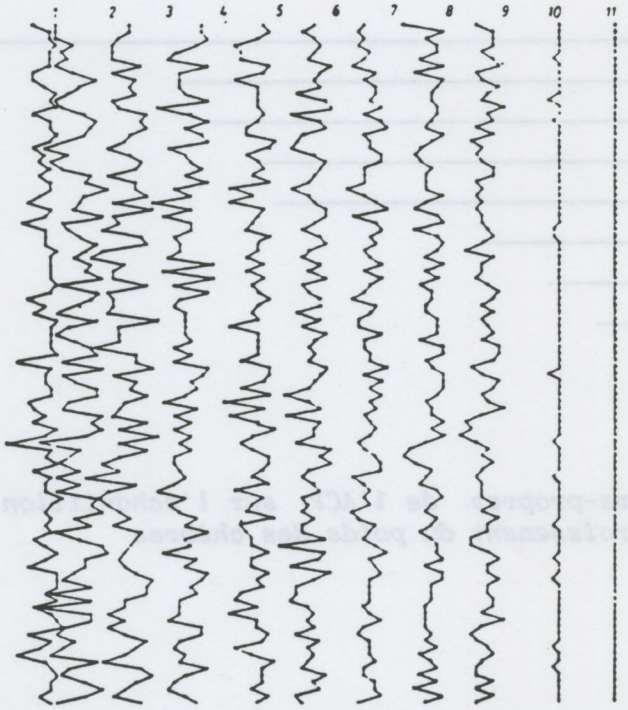


Fig. 5. Graphical representation of $A_k^{(b)}$ for 100 subsamples of the goat data. (Method [B]). The value of $A_k^{(b)}$ is given along the $x$-axes, the number of the subsample, $B$ is given in the $y$-axes.

### c/ Conclusions générales

• La méthode [B] surestime la variance de $\lambda_1$ par rapport à la méthode [A]. La différence entre les deux méthodes est particulièrement flagrante dans le premier exemple. De plus, pour cet exemple, l'estimation de $EQM_k$ obtenue avec la méthode [B] est supérieure à celle obtenue avec la méthode [A] lorsque $k = 1, 2$ et $3$.

Ces observations sont dues à l'approximation de la matrice des corrélations $R_x$ par la matrice de variance-covariance $V_w$, consentie dans l'utilisation de la méthode [B], basée sur le Bootstrap appliqué aux variables de l'ACP.

• On constate que $\bar{\lambda}_1^{(b)} \geqslant \lambda_1$ pour la méthode [A]. Ceci est en parfait accord avec les considérations théoriques exposées dans le chapitre 3. A l'inverse, le biais mesuré pour la dernière valeur-propre est négatif $\bar{\lambda}_p^{(b)} \leqslant \lambda_p$

• La règle empirique de détermination de la dimension du sous-espace de représentation en fonction de l'effondrement soudain des valeurs-propres est d'utilisation courante. Elle consiste à choisir le sous-espace de représentation k lorsqu'un écart important entre $\lambda_k$ et $\lambda_{k+1}$ est observé. Cette règle simple n'est pas vérifiée par les deux exemples étudiés.

Dans le premier exemple, elle conduirait à retenir les sous-espaces de dimension 1 ou 3. En effet, la proximité des valeurs de $\lambda_2$ et $\lambda_3$ conduirait à rejeter le sous-espace de dimension 2. La même considération est vraie pour $\lambda_4$ et $\lambda_5$. Or les estimations de la stabilité par la méthode du Bootstrap, telles que EQM, montrent que les quatre premiers axes sont stables.

Dans le second exemple, la même règle conduirait à choisir les sous-espaces de dimension 1 ou 5. Or ces deux sous-espaces sont très instables.

Le critère de choix du nombre de composantes basé sur "l'éboulis" des valeurs-propres (screegraph) ne peut donc, en aucun cas, se substituer à l'étude nécessaire de la stabilité d'une ACP.

• Enfin, la stabilité des derniers axes de l'ACP est souvent discutée. Dans le premier exemple, les premières variables de l'ACP sont stables tandis que les dernières sont instables. Au contraire, dans le second exemple, seules les dernières variables de l'ACP sont stables.

A l'instar du critère précédent, aucune règle basée sur le numéro des axes ne peut se substituer à l'étude de la stabilité d'une ACP, compte tenu des résultats obtenus.

# BIBLIOGRAPHIE SUR L'ETUDE DE LA STABILITE DE L'ACP


**Benasseni J.** (1985). *Influence des poids des unités statistiques sur les valeurs-propres en analyse en composantes principales.* Revue de Statistique Appliquée, XXXIII (4) : 41-55.

**Besse P., Ferre L.** (1993). *Sur l'usage de la validation croisée en analyse en composantes principales.* Revue de Statistique Appliquée, XLI (1) : 71-76.

**Daudin J.J., Duby C., Trecourt P.** (1988). *Stability of principal component analysis studied by the bootstrap method.* Statistics, 19 (2) : 241-258.

**Daudin J.J., Duby C., Trecourt P.** (1988). *PCA analysis studied by the bootstrap and the infinitesimal jackknife method.* Statistics, 20 (2) : 255-270.

**Krzanowski W.J.** (1987). *Cross-validation in principal component analysis.* Biometrics, 43 (3) : 575-584.